Polka Lines:
Learning Illumination Patterns and Reconstruction for Active Stereo
-Supplementary Document-

Seung-Hwan Baek Felix Heide
Princeton University

In this supplement, we present additional details and results. Specifically, we provide

- Description and results on the live capture system.
- Prototype DOE details.
- Optimization details on the DOE phase design.
- Radiometric calibration between DOE patterns.
- Results on the environment-specific training.
- Examples of the stereo NIR dataset.
- Additional analysis on the image formation model.
- Comparison between illumination patterns.
- Details on the self-supervised learning for finetuning.

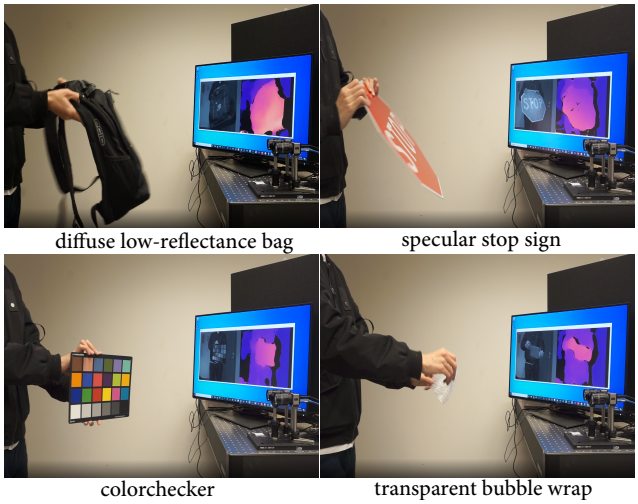

Figure 1. We demonstrate a live capture system from our polka-line prototype, reconstructing depth for several challenging objects in motion. Refer to the Supplemental Video.

1. Live Capture

We develop a live-capture system that acquires stereo images and estimates a disparity map at 10 frames per second (FPS) as shown in Figure 1. We refer to the Supplemental Video for the full results. We use a desktop computer with a NVIDIA GeForce RTX 3080 and the input 12-bit images are fed to our reconstruction network. We write our capture program in Python with multi threading to simultaneously perform capture and reconstruction, where Pytorch [4] is used for the reconstruction. The script consists of capturing the stereo images using the camera APIs, rectifying the images with the calibration data, and estimating a disparity map using our reconstruction network. For the reconstruction thread, we measure the elapsed time for each stage by averaging over 50 frames. It consists of four different stages: rectify, cpu-to-gpu transfer, depth reconstruction, and gpu-to-cpu transfer. Figure 3 shows the time took for each stage in the live-capture script.

Note that the current capture software is not fully optimized in terms of speed. Significant improvement could be obtained by using C++ implementation instead of the high-level Python API calls. Also, we expect employing the recent inference-dedicated network libraries such as NVIDIA TensorRT to reduce the processing time of the neural network. Another potential method is to downscale the input images while maintaining the depth reconstruction accuracy.

2. Prototype DOEs

We use a conventional photolithography process to prototype the three learned DOEs for indoor, outdoor, and general illumination conditions presented in the main paper. As we use the four-step lithography that produces 16 discrete height levels, we discretize the continuous height maps of the learned DOEs into the discrete forms. Figure 2 shows the simulated illumination patterns before and after the discretization process, demonstrating that the overall structured in the pattern remains same after the discretiza-

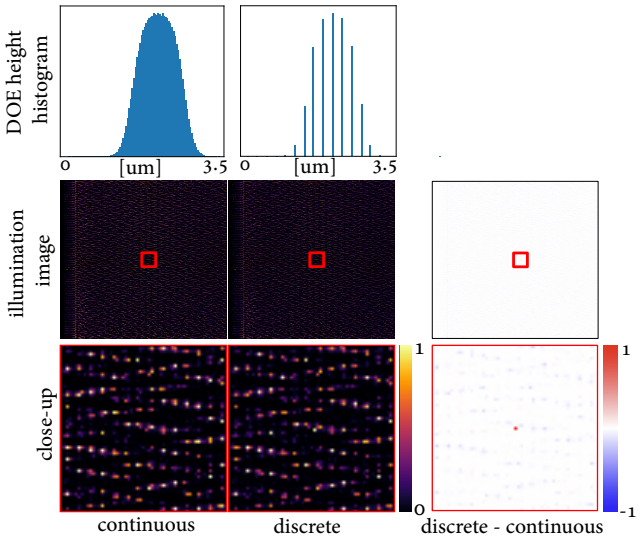

Figure 2. We discretize the optimized DOE height into 16 levels for photolithography fabrication. In simulation, the structure of the illumination image is maintained after the discretization process, except the amplified zero-mode diffraction.

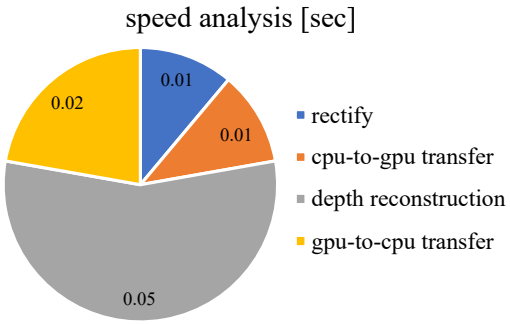

Figure 3. We measure the per-frame processing time and in their respective stage.

tion except the amplified zero-mode diffraction. This the zero-mode diffraction pattern is also observed in the illumination image of the fabricated DOEs shown in the main paper. These fabrication inaccuries can be mitigated in a commercial photolithography process where high-quality results from existing DOE-based structured-light systems have been presented as for the Intel D415 pattern that does not exhibit a zero order inaccuracy. We note that directly incorporating this discretization operator into our end-to-end learning framework induces training instability.

3. DOE Phase Design

Repurposing the proposed differentiable image formation model, we can design a DOE that produces a desired illumination pattern. We formulate this as an optimization problem of minimizing the difference between the target pattern image I_{target} and the simulated illumination image

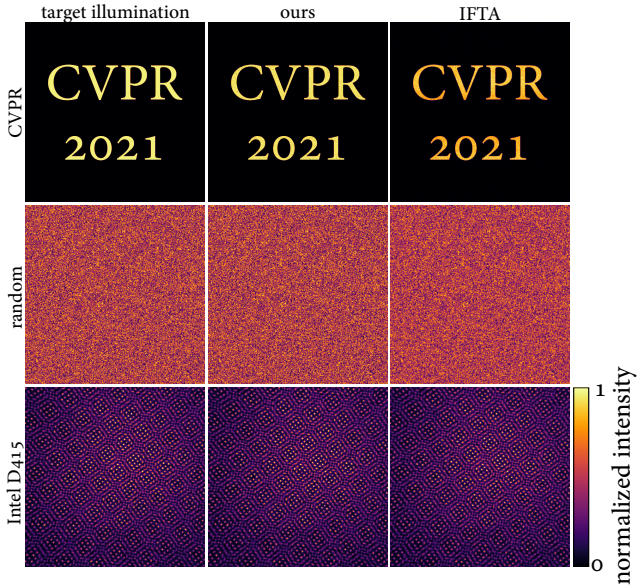

Figure 4. Our differentiable image formation can be used for designing a DOE that produces a desired illumination pattern. Our method performs outperforms the iterative Fourier transform method [1] with an advantage of design flexibility on the image formation and the loss function.

I_{illum} for a given phase map of the DOE ϕ as

$$\underset{\phi}{\text{minimize}}\ \mathsf{MSE}(I_{\text{illum}}(\phi), I_{\text{target}}), \tag{1}$$

where MSE is the mean squared error. As computing the illumination image $I_{\text{illum}}(\phi)$ consists of differentiable operations based on our image formation model, we can solve this problem relying on automatic differentiation using the Adam optimizer. Figure 4 shows target images and our reconstructions. We compare our method to the state-of-the-art iterative Fourier transform method [1] that indirectly solves the optimization problem. Our method not only outperforms this baseline in terms of reconstruction accuracy but also provides design flexibility by changing the image formation model and the loss function on demand.

4. Radiometric Calibration

In order to ensure fair comparison between different illumination patterns, we use the same illumination power across different patterns. In synthetic experiments, this is achieved by using the same parameter value of the laser power β. For the Intel D415 pattern, we obtain the power-normalized illumination pattern to apply the laser power β. To this end, we estimate the optimal illumination power β that reconstructs the captured Intel D415 pattern,

$$\underset{\beta,\phi}{\text{minimize}}\ \mathsf{MSE}(I_{\text{illum}}(\phi,\beta), I_{\text{target}}), \tag{2}$$

We use an integrating sphere of Thorlabs S142C and a power-controllable laser driver Thorlabs KLD101 to match the illumination power in for the prototype system.

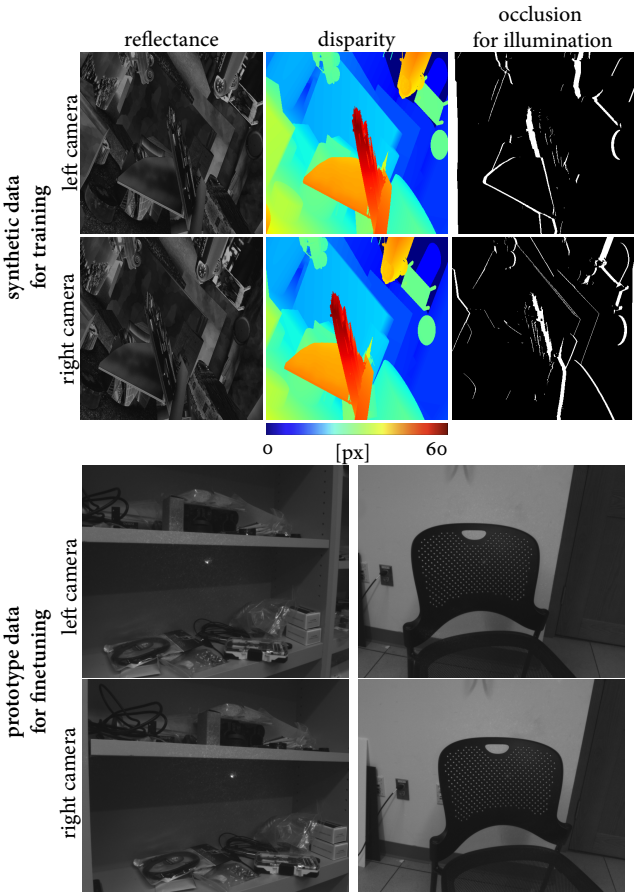

Figure 5. Our NIR-stereo datasets for the synthetic training and the finetuning.

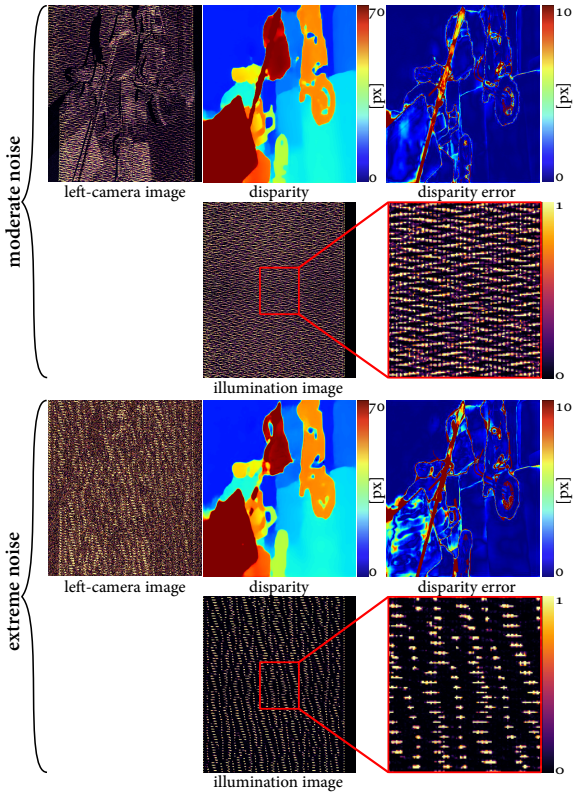

Figure 6. Optimized illumination and depth reconstruction for different noise levels. When the noise is moderate, the illumination pattern becomes dense with varying-intensity polka lines to provide dense correspondence cue. In contrast, severe noise makes the illumination pattern sparse with high intensities to stand out of the noise floor.

5. NIR-stereo Dataset

Our method use two different NIR-stereo datasets, each for training in simulation and finetuning for the experimental prototype. For the synthetic training, we modify the RGB-stereo dataset [3] as described in the main paper, resulting in 21718 training images and 110 testing images. For finetuning, we capture 76 real-world stereo images of indoor scenes. Figure 5 shows a sample in each dataset with varying reflectance and geometric complexity.

6. Prototype Calibration

We calibrate our experimental prototype for efficient stereo matching on the rectified domain. We capture a checkerboard at different positions and obtain the camera intrinsics, the distortion coefficients, and the extrinsic between the stereo cameras. The average reprojection error was 0.6 pixels. For each input stereo frame, we rectify the stereo images using the calibration data and feed them to the reconstruction network.

Then, we obtain the illumination images of the fabricated DOEs. For each DOE, we illuminate a white wall at a distance of 50 cm from the camera, while ensuring the intensity of the illumination pattern is in the observable dynamic range of the stereo cameras. We take the stereo images of the wall with and without the structured-light illumination. Using the no-illumination images as background, we compute the illumination images at the stereo viewpoints. Undistortion and rectification are applied to the illumination images in addition to the translational alignment with a manually-measured disparity of the wall. This procedure provides a high-quality illumination image at the rectified illumination viewpoint which can be used for the reconstruction network.

7. Environment-specific Illumination Design

Our method facilitates incorporating system and environment parameters in the image formation model, enabling us to design illumination patterns tailored to the given environments. Specifically, we evaluate the learned patterns in terms of ambient light and noise level.

Figure 6 shows the optimized illumination images and corresponding depth reconstructions for the moderate and the extreme noise levels. The standard deviations of the

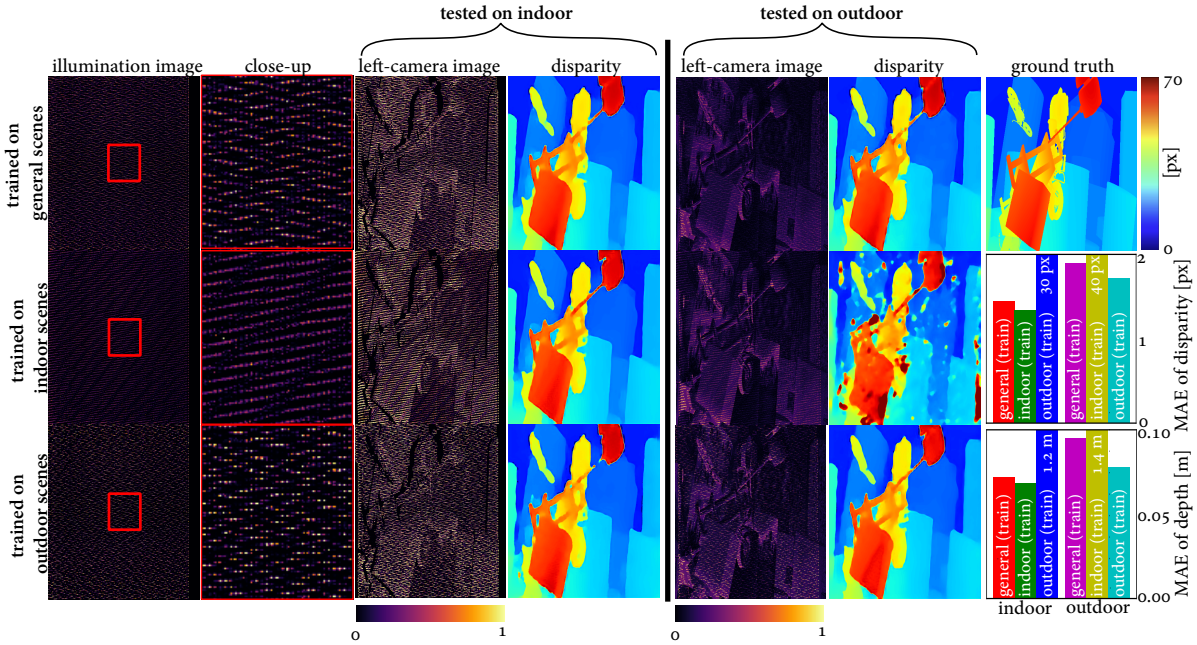

Figure 7. Our method enables us to design illumination patterns tailored for indoor, outdoor, or general environments.

Gaussian noise are 0.02 and 0.6 respectively. The extreme noise makes the illumination pattern sparse with high-intensity dots. This could be helpful to engrave the illumination features into the scene helping it stand out of the high-noise level. In the moderate noise case, we obtain dense varying-intensity polka lines in the illumination image, providing high-quality depth reconstructions.

We also test varying ambient-light power and laser power to simulate indoor and outdoor conditions by setting the parameter values of the ambient-light power and the laser power as follows: indoor ($\alpha = 0.0, \beta = 1.5$), outdoor ($\alpha = 0.5, \beta = 0.2$), and general ($\alpha \in [0, 0.5], \beta \in [0.2, 1.5]$). We train a DOE and a reconstruction network for each of the configurations. Figure 7 shows the optimized illumination patterns and their performance tested on both indoor and outdoor environments. We learn dense polka lines in the indoor scenes to provide many features for correspondence matching. For the outdoor scenes, we obtain sparse high-intensity polka lines, providing robustness against the strong ambient light and relatively weak laser power. Training on the general environment learns the polka lines with varying intensities with moderate density.

8. Illumination Patterns of Conventional DOEs

Our image formation model for an active stereo system includes computing the illumination image for a given DOE profile. As a sanity check on our image formation model, we compute the illumination patterns for two conventional DOE designs: random-height DOE and 2D diffraction grating. In theory, their illumination patterns are random dots and regular grid patterns with decaying intensity profile as the diffraction order increases. Figure 8 shows that our simulated illumination images produce such characteristics of the random dots and the regular grids.

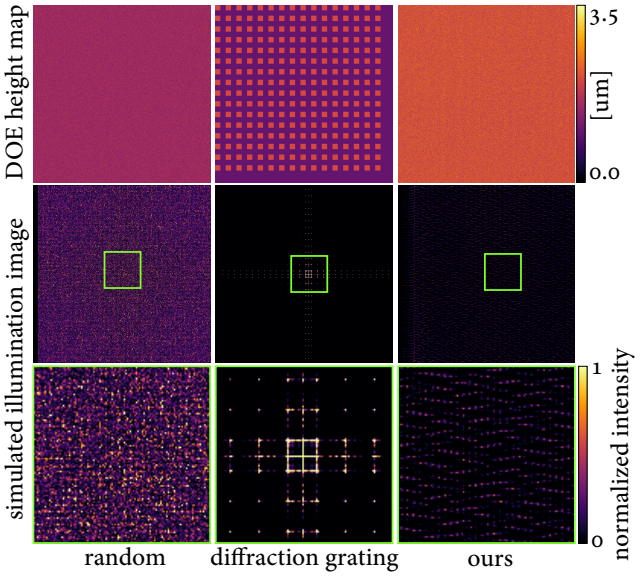

Figure 8. Our differentiable image formation can be applied to arbitrary DOE height maps, including random DOE height on the left, 2D diffraction grating on the middle, enabling the end-to-end design of illumination pattern for active-stereo systems.

9. Comparison with Illumination Patterns

We compare our learned polka-line pattern to the Intel D415 pattern and the ideal random-dots pattern in simula-

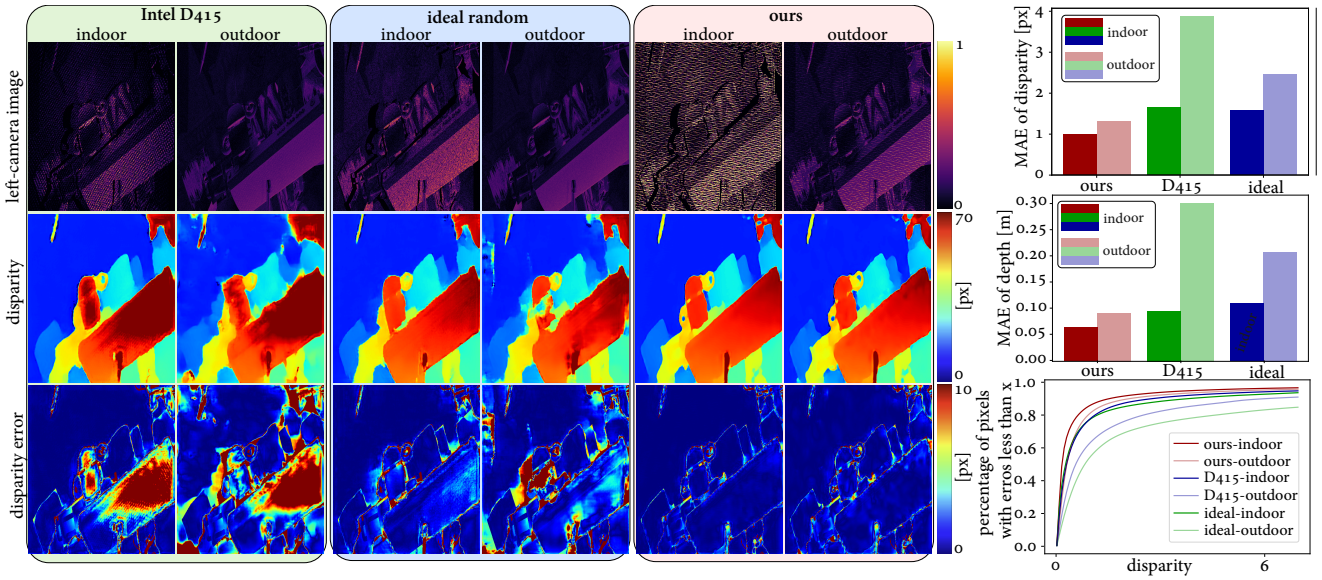

Figure 9. We compare the Intel D415 pattern, ideal random pattern, and our learned polka-line pattern in simulation. Our pattern outperforms the two hand-crafted illumination in all tested metrics.

tion. Figure 9 shows that the Intel D415 pattern suffers from sparse feature points, leading to reconstruction artifacts. The ideal random-dot pattern provides high-quality depth reconstruction on average, however degrades at the high ambient-light condition due to the scattered light energy by the random phase distribution. In contrast, our polka-line pattern provides accurate reconstruction with dense features and varying-intensity dots that we learn from the end-to-end optimization with the goal of accurate depth reconstruction.

Figure 10 shows the real-world comparison of the passive stereo, the Intel D415 pattern, and our polka-line pattern. Our polka-line design provides accurate reconstruction on feature-less objects with varying reflectance through the dense varying-intensity dots.

10. Self-supervised Finetuning

To handle the domain gap between the simulation and the real-world inputs, we finetune our reconstruction network. To this end, we first change the network architecture. Figure 11 shows the overview of the trinocular reconstruction network for finetuning. There are two major differences to the network used in simulation. First, we estimate disparity maps for both left and right views. This is implemented by computing the right-view disparity in a same way of computing the left-view disparity which is described in the main paper. Second, we introduce a validation network that estimates validity maps of the estimated disparity. Inspired by the left-right consistency [2], we warp the estimated left/right disparity maps to the other view and compute the difference with the original disparity maps. This difference and the stereo images are fed to the validation network as inputs. In summary, the changes of the network architecture and the loss function enables effective handling of challenging regions such as large occlusion and strong specularity which are often observed in the real-world inputs. Our finetuning is specifically formulated as the following optimization problem,

$$
\begin{aligned}
&\underset{\theta,\vartheta}{\text{minimize}}\, \mathcal{L}_u + \tau\mathcal{L}_v + \kappa\mathcal{L}_d, \\
&\mathcal{L}_u = \mathsf{MSE}\left(J^{L/R} \odot V_{\text{est}}^{L/R}(\vartheta), J_{\text{est}}^{L/R}(\theta) \odot V_{\text{est}}^{L/R}(\vartheta)\right), \\
&\mathcal{L}_v = \mathsf{CE}\left(V_{\text{est}}^{L/R}(\vartheta), \mathbf{1}\right), \\
&\mathcal{L}_d = l_2\left(\nabla D_{est}^{L/R}(\theta)\right),
\end{aligned}
\tag{3}
$$

where $V_{\text{est}}^{L/R}$ are the estimated left/right validity maps and $D_{\text{est}}^{L/R}$ are the corresponding disparity maps. $\mathcal{L}_u$ computes the mean squared error between the input and the estimated sensor images via validity-weighted warping: $J_{\text{est}}^{L/R} = \mathsf{warp}\left(J^{R/L}, D_{\text{est}}^{L/R}\right)$. $\mathcal{L}_v$ is the cross-entropy loss on the validity maps to avoid the trivial solution of making the validity as zero. $\mathcal{L}_d$ is the disparity smoothness loss to cope with real-world challenges in correspondence matching. τ and κ are the balancing weights set as 0.01 and 0.0001. The parameters of the reconstruction network θ are finetuned from the supervised training on the synthetic dataset, while the validation network parameters ϑ is trained from scratch. We train over 5 epochs for the finetuning.

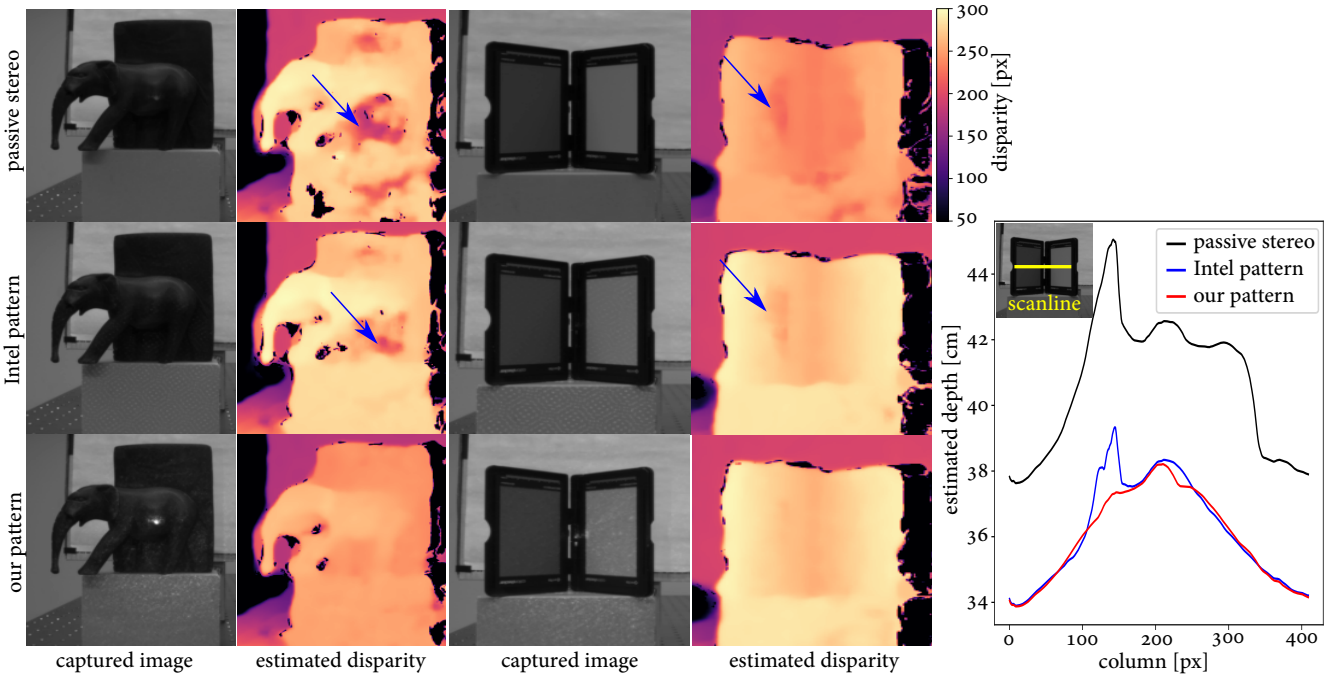

Figure 10. Our learned illumination pattern with varying-intensity dots outperforms passive stereo and the conventional fixed-intensity pattern (Intel D415 sensor) for high dynamic range of incident light. Blue arrows indicate estimation artifacts. We capture a v-shaped reflectance target (x-rite ColorChecker Pro Photo Kit) of which scanline analysis reveals the accurate reconstruction of the shape only by our polka-line pattern.

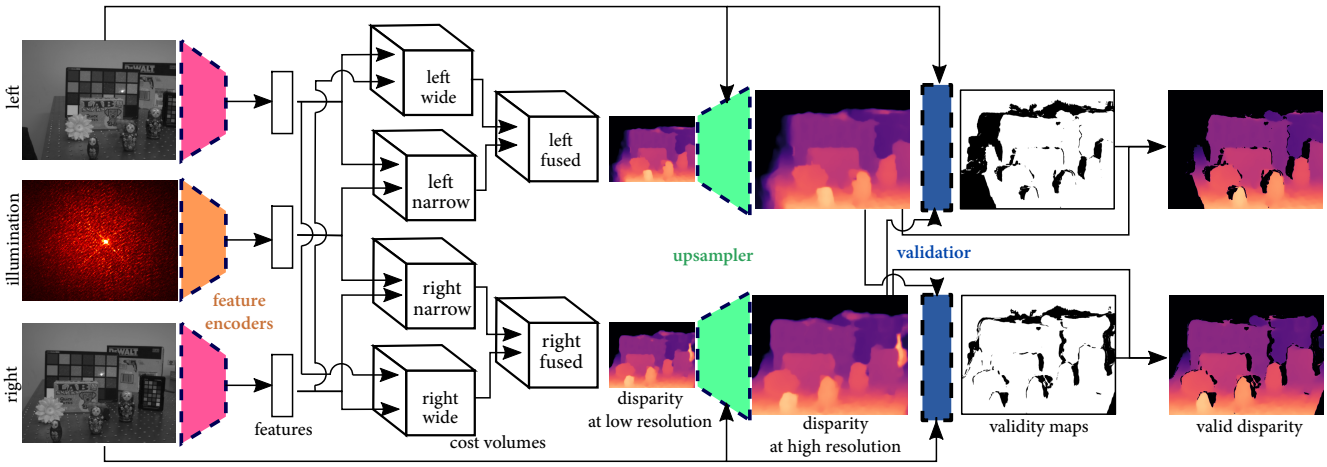

Figure 11. For finetuning, we extend the network architecture. We extract features for the left/right stereo images and the illumination image using the convolutional feature encoders. For each view, two cost volumes are constructed in narrow and wide baselines, fused into a multi-baseline cost volume. We estimate a disparity map for this cost-volume at a low spatial resolution which is upsampled to a original-resolution disparity using an edge-aware convolutional upsampler. The estimates disparity maps of the left and right views are then used for estimating validity maps that account for occlusion using a convolutional validator. The final disparity maps are obtained by making out the invalid region from the disparity estimates.